THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

CG dinucleotide clustering is a species-specific property of the genome

Citation for published version:
Glass, JL, Thompson, RF, Batbayar, K, Figueroa, ME, Olivier, EN, Oakley, EJ, Van Zant, G, Bouhassira, EE, Melnick, A, Golden, A, Fazzari, MJ & Greally, JM 2007, 'CG dinucleotide clustering is a species-specific property of the genome', *Nucleic Acids Research*, vol. 35, no. 20, pp. 6798-807. https://doi.org/10.1093/nar/gkm489

Digital Object Identifier (DOI):
10.1093/nar/gkm489

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Nucleic Acids Research

Publisher Rights Statement:
© 2007 The Author(s)

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 05. May. 2023

6798–6807 *Nucleic Acids Research, 2007, Vol. 35, No. 20* *Published online 10 October 2007*
doi:10.1093/nar/gkm489

CG dinucleotide clustering is a species-specific property of the genome

Jacob L. Glass[1], Reid F. Thompson[1], Batbayar Khulan[1], Maria E. Figueroa[2], Emmanuel N. Olivier[3], Erin J. Oakley[4], Gary Van Zant[4], Eric E. Bouhassira[3,5], Ari Melnick[2], Aaron Golden[6], Melissa J. Fazzari[7] and John M. Greally[1,5,*]

[1]Department of Molecular Genetics, [2]Department of Developmental and Molecular Biology and [3]Department of Cell Biology, Albert Einstein College of Medicine, Bronx, NY 10461, USA, [4]Division of Hematology/Oncology, University of Kentucky, Markey Cancer Center, 800 Rose Street, Lexington KY 40536, USA, [5]Department of Medicine (Hematology), Albert Einstein College of Medicine, Bronx, NY 10461, USA, [6]Department of Information Technology, National University of Ireland Galway, Newcastle Road, Galway, Republic of Ireland and [7]Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY 10461, USA

Received February 9, 2007; Revised May 23, 2007; Accepted June 6, 2007

ABSTRACT

Cytosines at cytosine-guanine (CG) dinucleotides are the near-exclusive target of DNA methyltransferases in mammalian genomes. Spontaneous deamination of methylcytosine to thymine makes methylated cytosines unusually susceptible to mutation and consequent depletion. The loci where CG dinucleotides remain relatively enriched, presumably due to their unmethylated status during the germ cell cycle, have been referred to as CpG islands. Currently, CpG islands are solely defined by base compositional criteria, allowing annotation of any sequenced genome. Using a novel bioinformatic approach, we show that CG clusters can be identified as an inherent property of genomic sequence without imposing a base compositional *a priori* assumption. We also show that the CG clusters co-localize in the human genome with hypomethylated loci and annotated transcription start sites to a greater extent than annotations produced by prior CpG island definitions. Moreover, this new approach allows CG clusters to be identified in a species-specific manner, revealing a degree of orthologous conservation that is not revealed by current base compositional approaches. Finally, our approach is able to identify methylating genomes (such as *Takifugu rubripes*) that lack CG clustering entirely, in which it is inappropriate to annotate CpG islands or CG clusters.

INTRODUCTION

Several observations have converged to focus attention on cytosine-guanine (CG) dinucleotide clusters in mammalian genomes. Digestion of genomic DNA with HpaII allowed the isolation of loci where these restriction sites cluster and are unmethylated in *cis*, defining a population of loci referred to as HpaII tiny fragments (1). Upon sequencing, these loci were found to be unusually rich in CG dinucleotides and (G + C) content when compared with other sequences in the 1985 Genbank database (2). The CpG island base compositional criteria now used for genomic annotation are derived from this original set of experiments.

With the appreciation that methylcytosine is unusually susceptible to mutation through deamination to thymine (3), a logical conclusion was that the absence of methylation at cytosines in CpG islands protected them from mutational decay during evolution. The implicit assumption is that these loci are universally unmethylated in normal cells but can be the target of abnormal methylation in cancer, or in unusual epigenetic regulatory processes such as genomic imprinting or X chromosome inactivation (4,5). CpG islands have proven valuable in focusing the study of the widespread genomic changes that occur in these processes, and are commonly used in designing custom microarrays for that purpose. CpG islands have also been used as a foundation for bioinformatic analyses such as finding gene promoters (6) and identifying sequence features that distinguish imprinted genes (7).

Despite its proven utility, there are problems with the original definition of CpG islands, including its lack of specificity. Using base compositional criteria alone, CpG island annotations identify over 350 000 sites in the

*To whom correspondence should be addressed. Tel: +1 718 430 2875; Fax: +1 718 824 3153; Email: jgreally@aecom.yu.edu

© 2007 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

human genome (Table 1), many of which are in repetitive sequences. Recognizing this problem, other groups have modified the original base compositional criteria (8) or the analytical approach used (9) in order to increase the stringency for CpG islands identification, greatly reducing the number of repetitive sequences annotated while preserving most CpG islands located at promoters. However, the more common approach, used by genome browsers such as that at UCSC (http://genome.ucsc.edu/) (10), is simply to remove all repetitive sequence prior to annotating CpG islands.

CpG island annotations are meant to identify constitutively unmethylated sites in the genome. However, the traditional CpG island criteria mostly identify repetitive sequences, as we show in Table 1, and these repeats are generally highly methylated (11). Furthermore, when we (12) and others (13) performed high-throughput cytosine methylation studies, even the annotated unique sequence CpG islands were subject to methylation at non-imprinted autosomal loci from normal tissues. It follows that the original base compositional criteria by themselves are not sufficient to predict methylation status.

Rather than modify existing base compositional criteria further, we decided to focus on the single characteristic of CG dinucleotides in which we had confidence: that they cluster at certain loci. We sought to identify whether such loci form a distinctive population within the genome as a whole. This approach allowed us to develop a new means of defining what we call CG clusters, and for the first time allows a species-specific definition that reveals the pattern of preservation of CGs to be genome specific and more conserved at orthologous loci than previously recognized. As CpG islands have been used as fundamental predictors of functionally important sites such as promoters (6), and we show that the CG cluster annotation has a substantially better positive predictive value for annotated transcription start sites than do CpG islands, it is likely that prior bioinformatic studies based on using CpG islands will be greatly improved by the use of CG clusters instead. We also show that the potential utility of CG clusters extends beyond sequence analysis alone, with demonstration of epigenetic predictive capacity, identifying substantially more hypomethylated sites than CpG islands in human CD34+ and embryonic stem cells. Because CpG islands are used as a basis for microarray studies of methylation changes, particularly in cancer (14), the use of CG clusters is likely to improve the sensitivity of such studies.

MATERIALS AND METHODS

CG cluster generation

The CG cluster annotation was generated using a set of custom PERL, R (http://www.r-project.org/) and shell scripts (available at http://greallylab.aecom.yu.edu/cgClusters/). Initially, the locations of every CG dinucleotide in the human genome were extracted from raw genomic DNA sequences (human May 2004 assembly hg17, http://genome.ucsc.edu/). Using these positions, every overlapping sequence fragment $S^n = \{S_1^n, S_2^n, \ldots\}$ containing a fixed number of CGs ($n = 5,10,\ldots 100$) and having variable length was identified. For each number n of CGs, the frequency of each fragment length was recorded and the distribution of fragment lengths was examined using the R statistical package for the presence of a short, CG-dense population $C^n = \{S_i^n | \text{length}(S_i^n) \leqslant \theta^n\}$ distinct from the longer fragments $C^{n\prime} = \{S_i^n | \text{length}(S_i^n) > \theta^n\}$. The threshold for each CG number θ^n (maximum fragment length) was defined to be the location of the local minimum in the fragment length histogram, estimated by identifying zero

Table 1. The total numbers of CpG islands and CG clusters in human and mouse, using unmasked sequence that contains repetitive elements, the UCSC CpG island track (generated from sequence in which transposons have been masked) and CG clusters excluding those with $\leqslant 27$ CGs derived from unique sequence (24 for mouse, the minimum number needed to define a CG cluster in each species).

		All	Sequence feature conserved at orthologous locus	Overlap with refSeq transcription start site
CpG islands (original definition[a])	Human[b]	350 201	42 445 (12.1%)	17 209 (4.9%)
	Mouse[c]	165 379	47 139 (28.5%)	11 822 (7.1%)
CpG islands (UCSC annotation[b])	Human	27 801	14 452 (52.0%)	14 121 (50.8%)
	Mouse	15 974	14 057 (88.0%)	9114 (57.1%)
CG clusters	Human	44 165	19 410 (44.0%)	16 822 (38.1%)
	Mouse	42 971	18 970 (44.2%)	11 859 (27.6%)
CG clusters (non-transposon)[d]	Human	31 225	19 071 (61.1%)	16 690 (53.5%)
	Mouse	21 587	17 614 (81.6%)	11 677 (54.1%)

The exclusion of transposon-derived CG clusters creates an annotation that is comparable to the UCSC CpG island annotations. We show the numbers and percentages of total for each sequence feature in terms of conservation of that sequence feature at the orthologous locus in the other species. We also quantify the numbers and proportions overlapping refSeq gene transcription start sites in each genome. Comparison of annotation performance in unmasked sequence shows CpG islands to suffer from excessive non-specificity, while the performance for non-repetitive sequences shows comparable proportional but quantitatively greater identification of conserved CG-dense regions or refSeq promoters using the CG cluster annotation in both human and mouse.

[a]Using cpgi130 program (4) (http:// http://cpgislands.usc.edu/) using parameters (G+C) $\geqslant$0.50, O/E CpG $\geqslant$0.60, window size $\geqslant$200 bp.
[b]Using hg17 assembly at the UCSC genome browser (http://genome.ucsc.edu/) (11).
[c]Using mm7 assembly at the UCSC genome browser.
[d]Removing CG clusters for which $\leqslant$27 CGs are contributed by unique sequence (24 CGs for mouse).

values of the first derivative of a cubic spline fit. Plots of θ^n against the number of CGs (n) exhibited a nearly linear relationship.

Mapping the CG-dense fragments in C^n back to the genomic sequence produces an annotation track where each annotated locus is a conglomeration of one or more overlapping fragments of variable length. However, the exact number, length and location of the annotated regions vary with the number of CGs per fragment (n). As the basis for choosing the optimal track in an objective manner, we noted that the fragments tended to aggregate and overlap to a greater extent in genomic regions of higher CG density. Because these types of regions are the major source of the CG-dense subpopulation, we used the number of overlapping fragments at locus j, ($|O_j^n|$), as a parameter for evaluating the information content of an annotated locus. To normalize for the length dependence of this value, we divided it by the maximum fragment length θ^n. To choose the track with maximal fragment overlap per locus, we compared genomic averages of this metric ($\sum_j |O_j^n|/\theta^n$) for different numbers of CGs per fragment (n). This allowed us to choose the species-specific optimal number of CGs per fragment for the final annotation. These annotations were then formatted for visualization in the UCSC genome browser and are available for download (human and mouse genomes) at http://greallylab.aecom.yu.edu/cgClusters/

Annotation track features including CpG islands and repetitive elements were examined using a local mirror of the UCSC genome browser MySQL database through the PERL DBI interface. The Takai and Jones (8) and Gardiner-Garden and Frommer (2) CpG island annotation tracks were generated using the cpgi130 program (8) (http://cpgislands.usc.edu/), and loaded into the database to facilitate analysis. The CG cluster annotation was also loaded into the database.

Analysis of CpG island and CG cluster promoter prediction was performed using a highly restrictive set of criteria. Only refSeq genes were considered, and promoter prediction was defined as strict overlap of the transcription start site. Non-transposon CG clusters were defined by quantifying the number of CG dinucleotides derived from transposon and unique sequences, identifying those for which unique sequence contributed less than the minimum number of CGs required for a CG cluster in each species and removing them from consideration. For the comparisons of CpG islands and CG clusters at orthologous promoters in human and mouse at the 23 loci, we used the same approach as in the original analysis (15), scoring conservation when the promoter of the gene had any overlap with the sequence feature. For the corresponding genome-wide analysis of CpG island and CG cluster conservation, we defined orthologous annotations in human and mouse using the mouse net (netMm7) track from the UCSC Genome Browser (16). Promoter hits were defined as strict overlap with transcription start sites of refSeq genes, while overlap of the annotation from one species with the annotation in the other species at orthologous sequences defined conservation of the CG-dense region.

Cytosine methylation analysis using the HELP (HpaII tiny fragment enrichment by ligation-mediated PCR) assay

Two normal human cell types were chosen for analysis, human embryonic stem cells and hematopoietic stem and progenitor cells. The H1 human embryonic stem cells (hESCs; NIH code WA01 from Wicell Research Institute, Madison, WI, USA) were cultured on P51R [hESC-derived MSCs (17)], plated at 75 000 cells per cm^2 or on matrigel (BD Biosciences, San Diego) at 37°C, 5% O_2 and 5% CO_2. The hESC medium contained DMEM/Ham's F-12, 20% Knockout Serum Replacer (KSR), 2 mM L-glutamine, minimal essential medium nonessential amino acid solution (NEAA), 0.1 mM penicillin–streptomycin 1% (all from Gibco, Grand Island, NY, USA), 4 ng/ml basic fibroblast growth factor (or 100 ng/ml for cells on matrigel, R&D Systems Inc., Minneapolis; or ProSpect-Tany, Technogene, Rehovot, Israel) and 0.1 mM 1-thioglycerol (Sigma–Aldrich, St Louis). The culture medium was changed daily, and the cells were passaged once weekly.

The hESCs were harvested using TrypLE™ EXPRESS (Gibco), washed and re-suspended in staining buffer [Dulbecco's phosphate-buffered saline (DPBS) + 5% KSR] at a concentration of 10^7 cells/ml and stained with mouse anti-human SSEA-4 antibody (DHSB) or isotype control (eBioscience, San Diego). Secondary staining was performed using rat anti-mouse IgG (H + L) immunoglobulin conjugated to fluorescein isothiocyanate (eBioscience). Based on fluorescence, positive cells for SSEA-4 were sorted using Moflow Cell-Sorter (DakoCytomation, Glostrup, Denmark). Genomic DNA was extracted from 1.5 to 2.5×10^6 cells using proteinase K digestion, phenol–chloroform extraction, dialysis against 0.2× SSC and concentration by surrounding the dialysis bag with PEG 20 000 to reduce water content by osmosis.

CD34+ cells were selected from bone marrow samples of healthy adult donors using a Miltyeni (Auburn, CA, USA) LS immunoabsorption column. Genomic DNA was extracted from 2 to 3×10^6 cells following a standard phenol–chloroform protocol followed by an ethanol precipitation and re-suspension of the DNA pellet in 10 mM Tris pH 8.0.

To identify hypomethylated loci, HELP analysis was performed (12) using a custom microarray representing HpaII-amplifiable sites at gene promoters (NimbleGen Systems). We used a categorical approach for the output of the assay, as our outcome of interest was defined in terms of methylated or hypomethylated loci. Methylated loci were identified by their inability to amplify from HpaII representations of genomic DNA (measured by the median microarray fluorescence intensities for the oligonucleotides representing each HpaII-amplifiable fragment, when median HpaII signal intensity was below the level of background signal, defined as 2.5 median absolute deviations above the median of random probe signal intensities), despite amplification in the corresponding MspI representation (signal intensity above the background calculated in the same way for the MspI channel).

The hypomethylated loci represented the remaining subset that was amplified in both channels (HpaII and MspI signal intensities above the levels of background signals). Overlap of each methylated or hypomethylated locus with CpG islands and CG clusters was quantified using a set of custom PERL scripts, and the results were analyzed by SQL query following their entry into a MySQL database.

RESULTS

We pursued the hypothesis that there is a subpopulation of sequences in the genome defined solely by their clustering of CG dinucleotides. This clustering is a result of the genome-wide decay of CG dinucleotide content, with preservation of CG density at certain regions. By measuring the distance spanned by a fixed number of CG dinucleotides for every such group genome-wide, we observed that there are two populations of loci with distinctive CG clustering densities (Figure 1a and b). Using the first local minimum in the distribution of spanned sequence fragment lengths as the boundary of the short, CG-dense population, we identified the maximum fragment length for each cluster corresponding to a fixed number of CGs. In analyzing these cutoffs, we defined a linear relationship between CG dinucleotide number and the associated maximum fragment length (Figure 1c).

The clear differentiation of CG-dense fragments from the rest of the genome provides a means of mathematically defining CG-dense regions and can therefore be used as a robust foundation for computational genomic annotation. Given a fixed number of CGs, the CG-dense fragments below the maximum fragment length could be identified and mapped back onto the genome. But, as Figures 2a and b show, each of the fixed number of CGs generates different annotations. Using fewer CGs and correspondingly smaller fragments, many small CG clusters are identified, whereas by using a greater number of CGs and correspondingly larger fragments, fewer clusters are identified, but each extends into large flanking regions of lower CG density.

We were able to optimize the criteria when we recognized that at any individual CG-dense locus, a given number of CGs generates multiple overlapping fragments. More CG-dense clusters require a greater number of fragments to span all of the CGs they contain. Accordingly, the more overlapping fragments that represent a given locus, the more likely it is to be significantly CG-dense. For each number of CGs, we calculated the number of overlapping fragments per cluster. We obtained a representation of information content for each CG number by summing this total across all loci in the genome and dividing by maximum fragment length. We then determined the optimal number of CGs per fragment using the maximum value obtained (Figure 2c). For the human genome, this optimum corresponds to 27 or more CG dinucleotides in a sequence of no more than 531 bp in length. This new means of identifying CG clusters is neither constrained by (G+C) content nor by the associated observed/expected CG dinucleotide ratio. In Figure 3, we show that the thresholds imposed by even the least stringent original base compositional criteria (2) cause many CG-dense loci in the genome to be missed. However, even though we are annotating the entire sequenced genome, including repetitive DNA, we identify only a small fraction of the ~350 000 CpG islands predicted by these old criteria (2) (Table 1).

We compared the functional significance of CpG islands and CG clusters in two ways—testing their relative

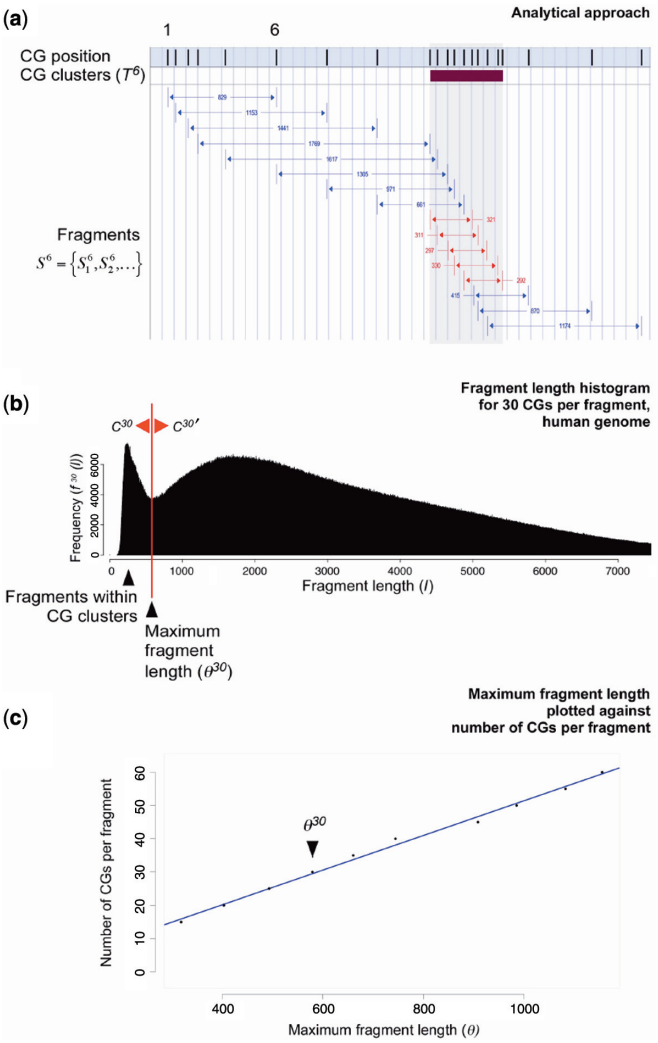

Figure 1. The analytical technique used to define CG clusters. First, a fixed number of CG dinucleotides is chosen, as illustrated in the example using six CGs (**a**). The first CG is identified in a chromosome, then the sixth, allowing the number of nucleotides between them to be recorded. The second and seventh CGs are then identified and the distance recorded and so on until the data for the entire genome is collected. When this analysis is performed for the entire human genome using 30 CGs at a time, the resulting lengths can be represented as a frequency histogram as shown in panel (**b**). Two populations are apparent—the peak on the left with short fragment lengths for this number of CGs, and the remainder of the genome where CGs do not cluster. The maximum fragment length for the clustered CGs is shown as a vertical red line. When the analysis is repeated for 5, 10, 15, ... 100 CGs genome-wide, different maximum fragment lengths for each number of CGs are derived, with a near-linear relationship between these variables as shown in panel (**c**). The arrowhead refers to the observation made for 30 CGs illustrated in panel (b).

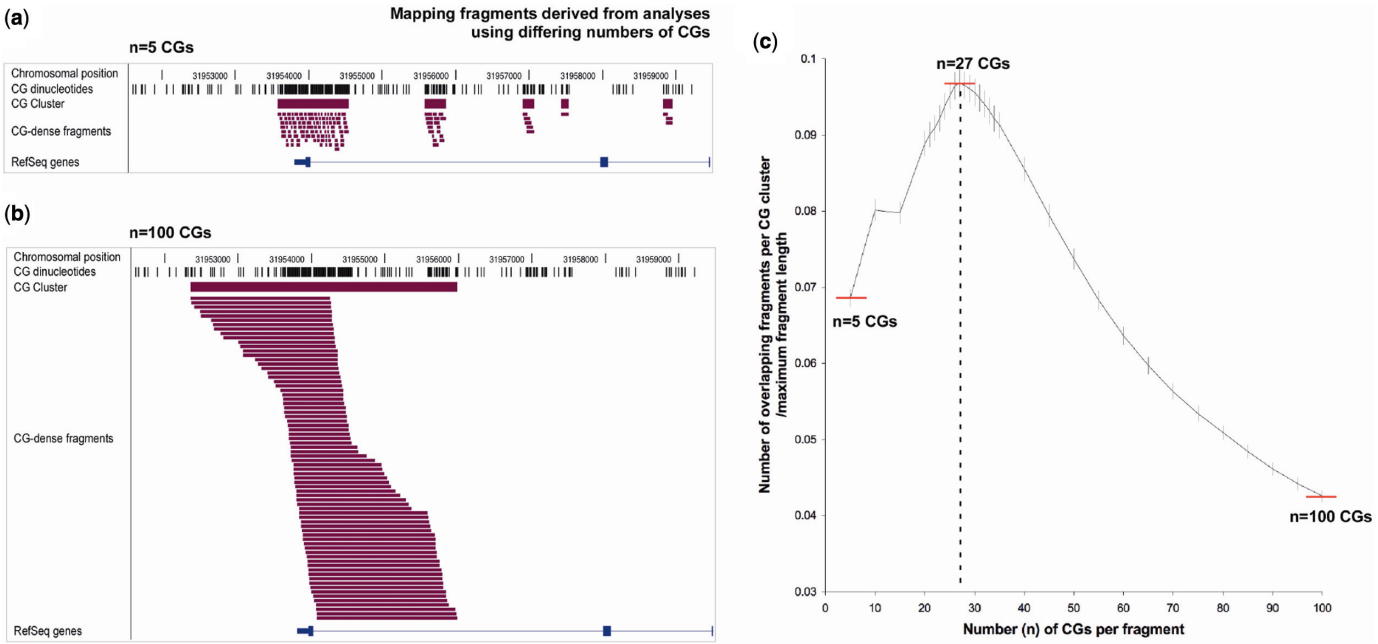

Figure 2. Creating a CG cluster annotation for the human genome. For a given number of CGs, significantly CG-dense fragments are defined as being shorter than the maximum fragment length. When these fragments are mapped back to the genome, some loci have multiple overlapping fragments, indicating that they are more likely to be CG-dense. These conglomerations define a genomic annotation track for each number of CGs used. Fewer CGs per fragment produces an annotation that is highly sensitive to local changes in CG density, defining a large number of small CG clusters, as shown in panel (**a**). On the other hand, a high number of CGs per fragment defines fewer CG clusters, but can extend far into flanking CG-poor regions (**b**). To find the intermediate optimum, we calculated the average number of fragments per CG cluster genome-wide. When recalculated relative to maximum fragment length, this measure of information content per CG cluster generated a peak at 27 CGs per fragment (**c**). This value is associated with a maximum fragment length value from the regression line in Figure 1c of 531 bp.

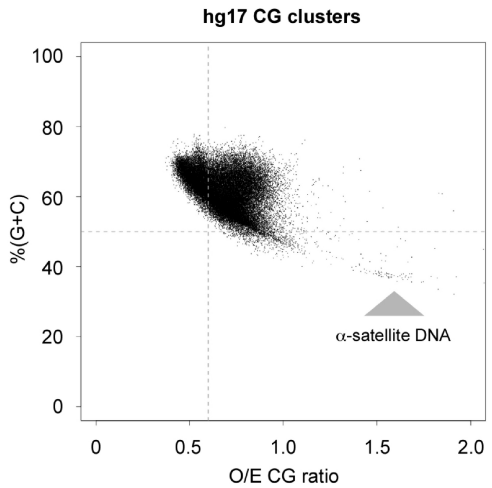

Figure 3. The base compositional characteristics of CG clusters (black) are shown in terms of observed to expected CG dinucleotide densities (O/E CG ratio) on the *x*-axis with (G + C) content on the *y*-axis. The dashed lines illustrate the relatively non-stringent thresholds of the original CpG island definition (3). Any points to the left of the vertical threshold or below the horizontal threshold show how many CG-dense loci would fail to be identified using base compositional criteria alone. The arrowhead illustrates extremely (A + T)-rich, CG-dense alpha satellite DNA sequences.

frequency co-localizing with promoters and with hypomethylated loci. A major use of the CpG island annotation has been to predict the location of transcription start sites in the genome. Approximately 40% (18) to 50% (6) of human promoters have been found to co-localize with CpG islands, while promoters of housekeeping genes have been described to have a near-universal association with CpG islands (18). We cross-correlated our CG clusters and the CpG island locations annotated at the UCSC Genome Browser with transcription start sites of refSeq genes from the same database, finding CpG islands to overlap 57% of refSeq transcription start sites, 79% of a published list of housekeeping genes (19) and 38% of a published list of tissue-specific genes (20). In contrast, the proportion of refSeq transcription start sites associated with CG clusters is substantially higher (68%, an additional 11% or 2701 refSeq transcription start sites), with 45% of the tissue-specific genes and 91% of housekeeping genes co-localizing with these CG clusters (Figure 4a; Table 1). As the UCSC CpG island annotation is of non-repetitive sequence and our CG cluster annotation was generated without this filter, we were concerned that the comparison was unfairly penalizing the UCSC CpG island annotation, so we tested the relative proportions of refSeq promoter overlaps for two other annotations, the 350 201 CpG islands that occur in the genome as a whole, and the 31 225 CG clusters that are not defined due to substantial overlap with transposable elements. In Figure 4b, we show that the performance of the CG cluster annotation is stronger for both unfiltered sequence (positive predictive values of 0.381 compared with 0.049) and non-transposon sequence (0.535 compared with 0.508) in identifying refSeq promoters. Similar patterns are found for the mouse genome (Table 1). Given that

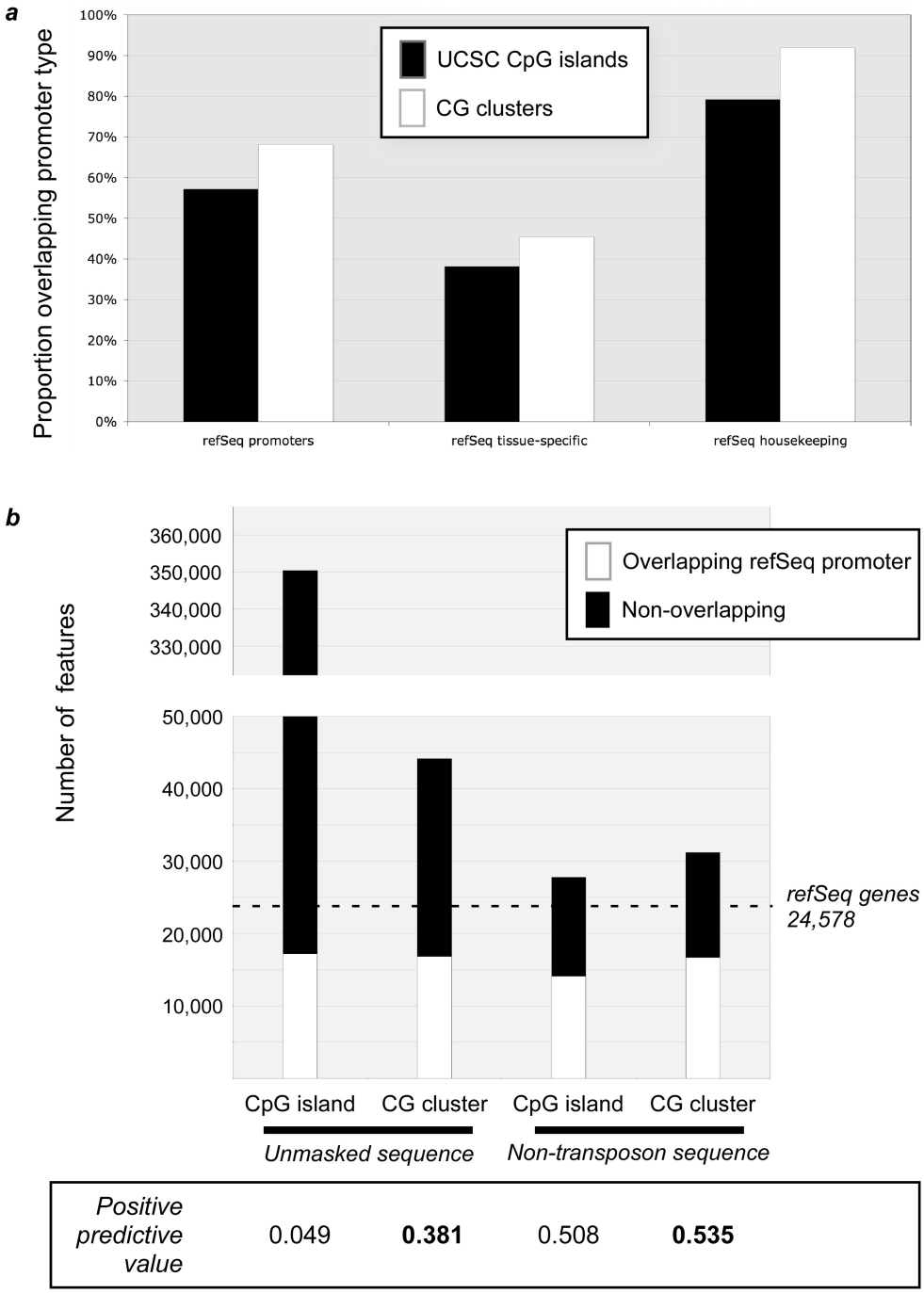

Figure 4. (**a**) CG clusters (white bars) overlap more refSeq transcription start sites than the CpG island annotations of the UCSC genome browser (black bars). CG clusters overlap the presumed promoters of a substantially higher proportion of genes overall (left), including both housekeeping and tissue-specific genes. In panel (**b**) we add in (in black) the number of loci that do not overlap refSeq promoters (white) to demonstrate the 'false positive' rate for the categories shown in panel (a), as well as the 350 201 CpG islands found when all genomic sequence is tested without removing repetitive elements and the CG clusters that are not defined by transposable elements. The positive predictive value for CpG islands and CG clusters (bold) are also shown, quantifying the relative performance of each of these annotations.

CpG islands have been used as a component of algorithms for predicting promoters in the genome (6), CG clusters should offer a more powerful resource for this and comparable purposes.

We tested the relative ability of CG clusters to detect hypomethylated sites by performing the HELP assay (12) on human embryonic stem cells and CD34+ hematopoietic stem and progenitor cells. A microarray representing HpaII-amplifiable fragments located near transcription start sites in the human genome was used for two biological replicates of each cell type. While similar proportions of loci at CpG islands and at CG clusters

demonstrated hypomethylation (Figure 5a), the absolute number of hypomethylated loci differed (Figure 5b), as the hypomethylated CpG islands represent a subset of the larger group of hypomethylated CG clusters. The CG clusters identify ~50% more hypomethylated loci than do CpG islands. We conclude that the CG cluster annotation is not only identifying more transcription start sites, it is also defining loci with comparable epigenetic characteristics.

We next addressed the question of why CpG islands are often not conserved between human and mouse (16,22). This is a puzzling issue if the CG-rich nature of the promoter is of functional importance, for example conferring the ubiquitous expression patterns that define housekeeping genes (18). It would be expected that such functional promoter characteristics would be conserved between species despite differences in overall CG dinucleotide content [observed/expected (O/E) CG ratios of 0.19 and 0.24 for mouse and human, respectively (22)]. It is therefore surprising that the total number of CpG islands in mouse is only ~58% of the number annotated for the human genome (Table 1).

When we performed the CG clustering analysis of the mouse genome, we found it also generates two populations with distinct CG density characteristics, but that the optimal CG cluster definition for the mouse genome is different from that of the human, corresponding to 24 or more CG dinucleotides in a sequence of no more than 585 bp in length (Figure 6). By comparison, human CG clusters consist of 27 CGs in no more than 571 bp. When we calculated the total number of CG clusters for the mouse genome, it was strikingly similar to that for the human (42 971 and 44 165, respectively, Table 1). In addition, when we re-analyzed a sample of 23 loci originally published to demonstrate the failure of CpG island conservation between these species (15), we found that while only 18 conserve CpG islands, 22 out of 23 conserve CG clusters, the single exception in this limited sample being the alpha globin orthologs (*HBA1*/*Hba-a1*). We extended this study to test conservation of each annotation genome-wide. Of all of the 27 801 CpG islands annotated at the UCSC Genome Browser, 14 452 have orthologous sequences with CpG islands in the mouse genome, while there exist 19 410 sites of conserved CG clustering (Table 1). When studied using our genome-specific annotations, clustered CG dinucleotides are demonstrably much more conserved between species than previously appreciated.

We extended the CG clustering histogram analysis to eight more genomes, including other organisms that are known to methylate their genomes, those that do so only transiently (*Drosophila melanogaster*) (23), and those that do not methylate at all. The surprising result of these analyses is that the fugu (Tiger Blowfish, *Takifugu rubripes*) genome, which has been described to methylate its DNA (24), does not exhibit uniquely CG-dense regions. What may explain this difference is that the degree of decay of CG dinucleotide content in the fugu genome is less than that of most genomes in which unique CG-dense regions emerge (Figure 7). The zebrafish (*Danio rerio*) genome, on the other hand, does display uniquely

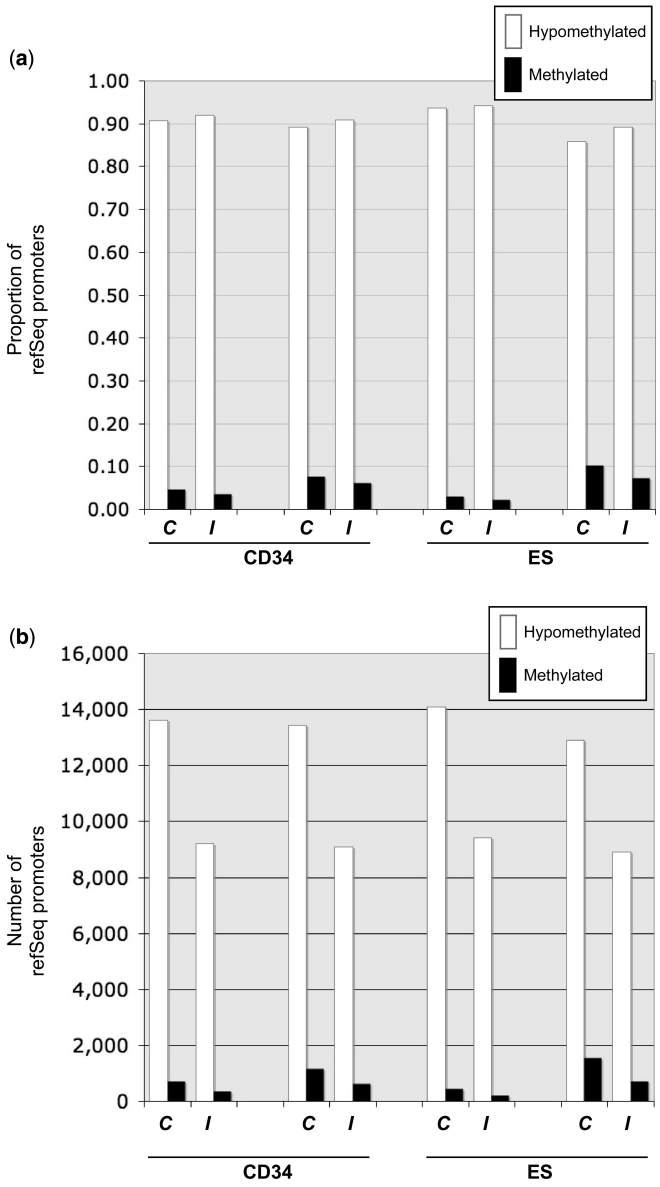

Figure 5. The HELP assay (13) was used on a custom promoter microarray to test cytosine methylation patterns in two samples each of CD34+ hematopoietic stem and progenitor cells (CD34) and human embryonic stem cells (ES). In panel (**a**) it is apparent that similar proportions of sites are categorized as hypomethylated for CG clusters (C) and CpG islands (I). However, the absolute number of hypomethylated sites detected by the CG cluster annotation is markedly larger than that for the CpG island annotation [panel (**b**)]. We conclude that the CG cluster annotation is not only detecting larger numbers of transcription start sites (Figure 4), it extends the ability of the CpG island annotation to identify more hypomethylated sites in the genome.

CG-dense regions with only marginally greater CG dinucleotide decay (O/E CG 0.53 as opposed to 0.57 in fugu). The remaining major difference between these genomes is that of size, the fugu genome being substantially smaller than the other methylating genomes at only 365 Mb total (25), a variable already suggested to be related to the evolution of cytosine methylation (26). Our data demonstrate that while cytosine methylation

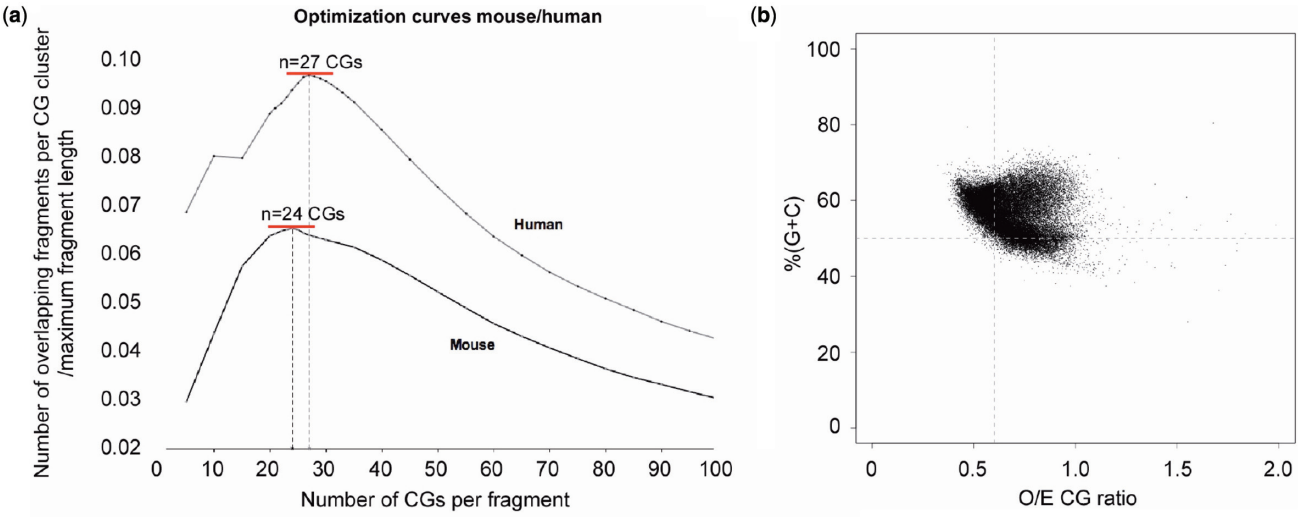

Figure 6. The mouse genome has different CG clustering characteristics than those of the human genome. The optimization curve characteristics for mouse are clearly different from those for human (**a**). The optimal mouse annotation contains fragments no longer than 585 nt with 24 or more CGs per fragment, fewer CGs in a longer stretch of DNA than for the human genome. In panel (**b**) it is again apparent that base composition criteria alone will fail to recognize a substantial proportion of CG-dense loci in this species.

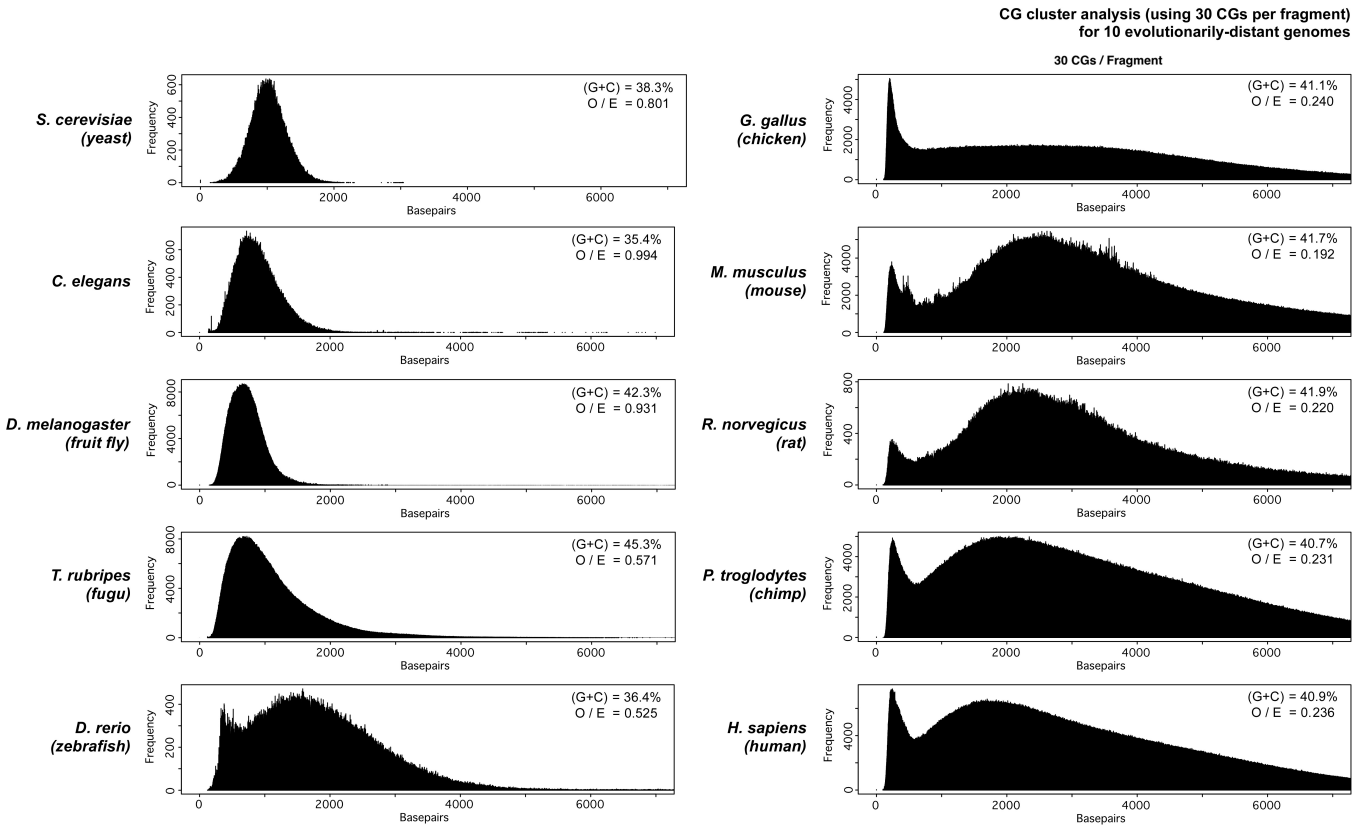

Figure 7. CG cluster analysis of 10 different species. These CG fragment length frequency plots were generated using 30 CGs per fragment for each species. Genomes containing CG clusters are defined by the distinct peak of short, uniquely CG-dense fragments. While the three non-methylating organisms on the left (*Saccharomyces cerevisiae*, *Caenorhabditis elegans* and *D. melanogaster*) show no uniquely CG-dense peak, it was surprising to find that fugu has similar characteristics despite the fact that it methylates its genome (25). Zebrafish, on the other hand, which also methylates its genome, has a distinct CG-dense peak, as do the other vertebrate genomes on the right. There is more CG decay in zebrafish than fugu (O/E CG ratios of 0.525 and 0.571, respectively), but this marginal difference does not appear sufficient to account for the emergence of CG-dense clusters in zebrafish. Methylation of the genome is not, therefore, always accompanied by the presence of CG-dense loci that avoid mutational decay. For a more detailed illustration of the CG cluster analysis of these genomes, see the Supplementary Data section.

appears to be necessary for CG decay, it is not sufficient to cause local preservation of clustered CG dinucleotides. Furthermore, we can conclude that any annotation of the fugu genome to indicate the presence of CpG islands or CG clusters is inappropriate.

DISCUSSION

The approach of testing for CG clustering reveals the loci at which CG dinucleotide decay occurred at a markedly lower rate than it did in the rest of the genome. Unlike CpG islands, CG clusters occur with a large range of (G + C) content and O/E CG ratios, revealing the unusual CG density of loci such as alpha satellite DNA sequences, targets for the DNMT3B methyltransferase (27) that are very (A + T)-rich and are consequently not defined as distinctive using traditional CpG island base compositional criteria. While this is an example of the CpG island definition being excessively restrictive, the definition also suffers from the problem of identifying a large number of sites located within transposable elements. It is certainly possible that these retroelements may have *cis*-regulatory effects (28), but their overall tendency to be methylated (11) diminishes the usefulness of the CpG island annotation as a mark of unmethylated DNA in the genome, a major use of the annotation in cancer epigenomics, especially in defining the CpG island methylator phenotype (CIMP), which requires that methylation of a given CpG island be distinctive in neoplastic cells (29). The approach used in the creation of CpG island browser tracks involves using sequences from which repetitive elements have previously been removed, increasing the likelihood of identifying constitutively unmethylated, presumably *cis*-regulatory sequences, but even this approach defines a number of sites that are methylated in normal cells (12,13).

Our HELP data indicate that CG clusters perform better than CpG islands at identifying unmethylated sites in the genome. However, our data do not support CG clusters being universally unmethylated, as we find satellite DNA and young retrotransposons to encode CG clusters, and these sequences are normally methylated (11,27). We propose that what distinguishes clustered from non-clustered CGs in the genome is the greater stability of associated epigenetic marks, such as hypomethylation at gene promoters or methylation of alpha satellite DNA.

CpG islands have recently been described to be located at the bivalent domains of histone tail modifications in embryonic stem cells (30), reinforcing the rationale for using these loci as a means of identifying candidate *cis*-regulatory sites in the genome. CpG islands represent a foundation annotation of the genome on which other annotations are built, for example contributing to algorithms to identify gene promoters (31). However, we show (Figure 3) that the CG cluster annotation performs substantially better than the CpG island annotation in localizing to known promoters (as represented by refSeq transcription start sites), indicating that an improved foundation annotation like CG clusters may improve the performance of algorithms currently using CpG islands.

Because identical criteria have been used to define CpG islands in different species, in which CG clustering can have markedly different characteristics, CpG islands have been thought to be poorly conserved between species, especially with the focus on human/mouse comparisons (21). We show that a species-specific definition of CG clusters reveals an unexpected degree of conservation of this annotation between human and mouse. We anticipate that conserved CG clusters will represent a subset of loci of exceptional functional importance in the genome. However, it is also clear (Figure 7) that it is inappropriate to annotate all methylating genomes for the presence of CG-dense regions. Fugu is annotated at genome browsers (UCSC and Ensembl) for the presence of CpG islands despite the fact that it does not have a distinctive population of loci maintaining CG content in an overall genomic context of CG decay. Zebrafish, on the other hand, with a similar degree of CG decay, manifests the two populations of CG content. Interestingly, the methylation of cytosines in the zebrafish genome includes a substantial proportion at non-CG dinucleotide sites (32), yet the selective preservation of CG content at a subset of loci occurs in this genome.

CONCLUSIONS

We show that CG clusters, when present in a genome, define themselves as a distinctive population of loci. This novel annotation performs better at identifying promoters and hypomethylated DNA than current CpG island definitions, and allows a species-specific definition of CG clusters that reveals a previously unsuspected degree of conservation of this sequence feature. The species specificity of what defines a CG cluster indicates that CG dinucleotides only need to be enriched within the context of their genome to be distinctive and presumably functional. We expect that the annotations of CG clusters will prove valuable to those studying the genome as well as the epigenome, so we have provided the human and mouse annotations as a resource for public use at http://greallylab.aecom.yu.edu/cgClusters/

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

URLS

The UCSC Genome Browser (http://genome.ucsc.edu/), the ENSEMBL Genome Browser (http://www.ensembl.org/), the Greally lab CG cluster webpage (http://greallylab.aecom.yu.edu/cgClusters/), the R project (http://www.r-project.org/) and the CpG island searcher website with the cpgi130 program (http://cpgislands.usc.edu/).

ACKNOWLEDGEMENTS

The following shared resources at Albert Einstein College of Medicine were used to generate data for this study: the Bioinformatics Shared Resource, the Genomics Core Facility, as well as resources from the Albert Einstein Cancer Center. This work is supported by a grant from the National Institutes of Health (NIH; NICHD) R01 HD044078 to J.M.G. Both J.L.G. and R.F.T. are supported by NIH MSTP Training Grant GM007288. A.G. is supported by Science Foundation Ireland (Grant Number 05/RFP/CMS0001). G.V.Z. is supported by NIH grants R01 AG022859 and R01 AG024950. The authors thank Ms Carol Swiderski for her expert technical assistance. Funding to pay the Open Access publication charges for this article was provided by R01 HD044078 from the NIH.

Conflict of interest statement. None declared.

REFERENCES

1. Bird,A.P. (1986) CpG-rich islands and the function of DNA methylation. *Nature*, **321**, 209–213.
2. Gardiner-Garden,M. and Frommer,M. (1987) CpG islands in vertebrate genomes. *J. Mol. Biol.*, **196**, 261–282.
3. Duncan,B.K. and Miller,J.H. (1980) Mutagenic deamination of cytosine residues in DNA. *Nature*, **287**, 560–561.
4. Li,E., Beard,C. and Jaenisch,R. (1993) Role for DNA methylation in genomic imprinting. *Nature*, **366**, 362–365.
5. Pfeifer,G.P., Tanguay,R.L., Steigerwald,S.D. and Riggs,A.D. (1990) In vivo footprint and methylation analysis by PCR-aided genomic sequencing: comparison of active and inactive X chromosomal DNA at the CpG island and promoter of human PGK-1. *Genes Dev.*, **4**, 1277–1287.
6. Ioshikhes,I.P. and Zhang,M.Q. (2000) Large-scale human promoter mapping using CpG islands. *Nat. Genet.*, **26**, 61–63.
7. Luedi,P.P., Hartemink,A.J. and Jirtle,R.L. (2005) Genome-wide prediction of imprinted murine genes. *Genome Res.*, **15**, 875–884.
8. Takai, D. and Jones, P.A. (2002) Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc. Natl. Acad. Sci. USA*, **99,** 3740–3745.
9. Hackenberg,M., Previti,C., Luque-Escamilla,P.L., Carpena,P., Martinez-Aroza,J. and Oliver,J.L. (2006) CpGcluster: a distance-based algorithm for CpG-island detection. *BMC Bioinformatics*, **7**, 446.
10. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
11. Yoder,J.A., Walsh,C.P. and Bestor,T.H. (1997) Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet.*, **13**, 335–340.
12. Khulan,B., Thompson,R.F., Ye,K., Fazzari,M.J., Suzuki,M., Stasiek,E., Figueroa,M.E., Glass,J.L., Chen,Q. *et al.* (2006) Comparative isoschizomer profiling of cytosine methylation: the HELP assay. *Genome Res.*, **16**, 1046–1055.
13. Song,F., Smith,J.F., Kimura,M.T., Morrow,A.D., Matsuyama,T., Nagase,H. and Held,W.A. (2005) Association of tissue-specific differentially methylated regions (TDMs) with differential gene expression. *Proc. Natl Acad. Sci. USA*, **102**, 3336–3341.
14. Issa,J.P. (2000) CpG-island methylation in aging and cancer. *Curr. Top. Microbiol. Immunol.*, **249**, 101–118.
15. Matsuo,K., Clay,O., Takahashi,T., Silke,J. and Schaffner,W. (1993) Evidence for erosion of mouse CpG islands during mammalian evolution. *Somat. Cell. Mol. Genet.*, **19**, 543–555.
16. Kent,W.J. (2002) BLAT – the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
17. Olivier,E.N., Rybicki,A.C. and Bouhassira,E.E. (2006) Differentiation of human embryonic stem cells into bipotent mesenchymal stem cells. *Stem Cells*, **24**, 1914–1922.
18. Larsen,F., Gundersen,G., Lopez,R. and Prydz,H. (1992) CpG islands as gene markers in the human genome. *Genomics*, **13**, 1095–1107.
19. Eisenberg,E. and Levanon,E.Y. (2003) Human housekeeping genes are compact. *Trends Genet.*, **19**, 362–365.
20. Zhang,L. and Li,W.H. (2004) Mammalian housekeeping genes evolve more slowly than tissue-specific genes. *Mol. Biol. Evol.*, **21**, 236–239.
21. Antequera,F. and Bird,A. (1993) Number of CpG islands and genes in human and mouse. *Proc. Natl Acad. Sci. USA*, **90**, 11995–11999.
22. Fazzari,M.J. and Greally,J.M. (2004) Epigenomics: beyond CpG islands. *Nat. Rev. Genet.*, **5**, 446–455.
23. Lyko,F., Ramsahoye,B.H. and Jaenisch,R. (2000) DNA methylation in Drosophila melanogaster. *Nature*, **408**, 538–540.
24. Armes,N., Gilley,J. and Fried,M. (1997) The comparative genomic structure and sequence of the surfeit gene homologs in the puffer fish Fugu rubripes and their association with CpG-rich islands. *Genome Res.*, **7**, 1138–1152.
25. Aparicio,S., Chapman,J., Stupka,E., Putnam,N., Chia,J.M., Dehal,P., Christoffels,A., Rash,S., Hoon,S. *et al.* (2002) Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science*, **297**, 1301–1310.
26. Rollins,R.A., Haghighi,F., Edwards,J.R., Das,R., Zhang,M.Q., Ju,J. and Bestor,T.H. (2006) Large-scale structure of genomic methylation patterns. *Genome Res.*, **16**, 157–163.
27. Hassan,K.M., Norwood,T., Gimelli,G., Gartler,S.M. and Hansen,R.S. (2001) Satellite 2 methylation patterns in normal and ICF syndrome cells and association of hypomethylation with advanced replication. *Hum. Genet.*, **109**, 452–462.
28. Greally,J.M. (2002) Short interspersed transposable elements (SINEs) are excluded from imprinted regions in the human genome. *Proc. Natl Acad. Sci. USA*, **99**, 327–332.
29. Issa,J.P. (2004) CpG island methylator phenotype in cancer. *Nat. Rev. Cancer*, **4**, 988–993.
30. Bernstein,B.E., Mikkelsen,T.S., Xie,X., Kamal,M., Huebert,D.J., Cuff,J., Fry,B., Meissner,A., Wernig,M. *et al.* (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, **125**, 315–326.
31. Davuluri,R.V., Grosse,I. and Zhang,M.Q. (2001) Computational identification of promoters and first exons in the human genome. *Nat. Genet.*, **29**, 412–417.
32. Shimoda,N., Yamakoshi,K., Miyake,A. and Takeda,H. (2005) Identification of a gene required for de novo DNA methylation of the zebrafish no tail gene. *Dev. Dyn.*, **233**, 1509–1516.